Claims

1. A process for testing an evaluation data record having attribute fields containing data comprising:

providing a reference table having a number of reference records against which a evaluation data record is tested;

identifying reference table tokens contained within the reference records of the reference table and determining a count of tokens in the reference table classified according to attribute field; and

assigning a similarity score to said evaluation data record in relation to a reference record within the reference table based on a combination of:

the number of common tokens of an evaluation field of the input data record and a corresponding field within a reference record from the reference table;

the similarity of the tokens that are not the same in the evaluation field of the input data record and the corresponding field of the reference record from the reference table; and

a weight of the tokens of the evaluation data record that is based on a count of the tokens from a corresponding field contained within the reference table.

2. The process of claim 1 wherein a look-up table based of contents of reference records in the reference table is prepared before evaluation of the evaluation data record and wherein the tokens of the evaluation data record are evaluated by comparing the contents of the look-up table with contents of the tokens of said evaluation data record to prepare a candidate set of reference records for which a similarity score is assigned.

3. The process of claim 2 additionally comprising a step of evaluating tokens in the reference table by:

breaking tokens in the reference table up into sets of substrings having a length q;

applying a function to the set of substrings for a token to provide a vector representative of a token; and

building a lookup table for substrings found within the tokens that make up the reference table.

4. The process of claim 3 wherein the process of building the lookup table creates an entry for each substring comprising: an attribute field for said substring, a co-ordinate within a vector for said substring, a frequency of said substring, and a list of reference records where said substring appears in the specified attribute field and vector co-ordinate position.

5. The process of claim 4 wherein the weights that are assigned to tokens of the evaluation record are distributed across candidate records from the reference table during a determination of a candidate set of records.

6. The process of claim 4 wherein a candidate record table is built and records listed in the lookup table are added to the candidate record table based on vector representations of the tokens of the input record.

7. The process of claim 6 wherien a candidate record is added to the candidate record table only if a score assigned to the reference record can exceed a threshold based on an already evaluated substring.

8. The process of claim 6 wherein once a likely reference record that matches the evaluation data record with a specified degree of certainty is found, further searching for records in the reference table is stopped.

9. The process of claim 1 wherein a closest K reference records from the reference table are identified as possible matches with the input record.

10. The process of claim 1 wherein reference records having a similarity score greater than a threshold are identified as candidate records.

11. The process of claim 2 additionally comprising a step of evaluating tokens in the reference table by applying a function to the set of substrings for a token to provide a vector representative of a token; and further comprising preparing the look-up table for tokens that make up the reference table by creating an entry in the look-up table for a token including an attribute field for the token or a substring, an attribute field for a co-ordinate within a vector for said token or substring, an attribute field for a frequency of said token or substring, and a list of reference records where said token or said substring appears in the specified field and vector co-ordinate position.

12. The process of claim 1 additionally comprising the step of maintaining a token frequency cache in a high speed access memory for use in assigning weights to said tokens.

13. The process of claim 1 wherein the tokens in different attribute fields are assigned different weights in determining said score.

14. The process of claim 1 wherein assigning a score includes determining a cost in transposing the order of two tokens in determining a similarity between tokens of the input data record and records in the reference table.

15. The process of claim 14 wherein the determining of a cost in transposing tokens takes into account a weight of said tokens that are transposed.

16. A system for evaluating an input data record having fields containing data comprising:

a database for storing a reference table having a number of records against which an input data record is evaluated;

a preprocessor component for evaluating records in the reference table to identify tokens and determining a count of tokens in the reference table classified according to record field; and

a matching component for assigning a score to an input data record in relation to a reference record within the reference table based on a combination of:

i) the number of common tokens of an evaluation field of the input data record and a corresponding field within a reference record from the reference table;

ii) the similarity of the tokens that are not the same in the evaluation field of the input data record and the corresponding field of the reference record from the reference table; and

iii) a weight of the tokens of the evaluation data record that is based on a count of the tokens from the corresponding field contained within the reference table.

17. The system of claim 16 wherein the preprocessor component evaluates tokens in the reference table by:

breaking tokens in the reference table up into sets of substrings having a length q;

applying a hash function to the set of substrings for a token to provide a vector representative of a token; and

building a lookup table for substrings found within the tokens that make up the reference table.

18. The system of claim 17 wherein the preprocessor creates an entry in the lookup table for each substring, an attribute field for said substring, a co-ordinate within a vector for said substring, and a list of reference records where said substring appears in the specified attribute field and vector co-ordinate position.

19. A process for evaluating an input data record having attribute fields containing data comprising:

providing a number of reference records organized into attribute fields against which an input data record is evaluated;

evaluating reference records to identify tokens from said attribute fields and then evaluating each token to build a vector of token substrings that represent the token;

building an index table wherein entries of the index table contains a token substring and a list of reference records that contain a token that maps to the token substring; and

looking up reference records in the index table based on the contents of the input record and selecting a number of candidate records from the reference records in the index table for comparing to said input data record.

20. The process of claim 19 additionally comprising a step of assigning a similarity score to said input data record in relation to a candidate set of reference records based on a combination of:

the number of common tokens of an evaluation field of the input data record and a corresponding field within a reference record;

the similarity of the tokens that are not the same in the evaluation field of the input data record and the corresponding field of the reference record; and

a weight of the tokens in the corresponding field of said reference records based on a count of the tokens from the corresponding field contained within the reference records.

21. The process of claim 19 wherein a candidate record table is built and candidate records from the index table are added to a candidate record table based on an H dimensional vector of token substrings determined from tokens contained in the input record.

22. The process of claim 21 wherein tokens are parsed from the input data record and tokens contained in said input data record are assigned token weights based on occurrences of the tokens in the reference table and further wherein records added to the candidate record table are factored by an amount corresponding to the weights of tokens extracted from the input data record.

23. The process of claim 22 wherein weights are assigned to tokens based on the attribute field in which the tokens are contained in the reference table.

24. The process of claim 19 additionally comprising a step of assigning a similarity score to said input data record in relation to a candidate set of reference records based on :

a cost in converting tokens in the input data record to tokens in a corresponding field of a reference record wherein the cost is based on a weight of the tokens in the corresponding field of said reference record corresponding to a count of the tokens from the corresponding field contained within the reference records.

25. The process of claim 19 wherein the reference records are stored in a reference table and wherein a candidate record table is built and candidate records from the index table are added to a candidate record table based on token substrings contained in the input record and wherein tokens contained in said input data record are assigned token weights based on occurrences of the tokens in the reference table and further wherein records added to the candidate record table are factored by an amount corresponding to the weights of tokens contained in the input data record.

26. The process of claim 21 wherein a candidate record is added to the candidate record table only if a possible score assigned to the reference record in the reference table can exceed a threshold based on an already evaluated substring.

27. The process of claim 26 wherein once a likely reference record that matches the evaluation data record with a specified degree of certainty is found further searching for reference records in the reference table is stopped.

28. The process of claim 20 wherein a closest K reference records from the reference table are identified as possible matches with the input record.

29. The process of claim 20 wherein reference records having a similarity score greater than a threshold are identified as candidate records.

30. The process of claim 20 additionally comprising the step of maintaining a token frequency cache in a high speed access memory for use in assigning weights to said tokens.

31. The process of claim 20 wherein the tokens in different attribute fields are assigned different weights in determining said score.

32. The process of claim 19 wherein the index table additionally comprises an attribute field for a token from which a substring is derived.

33. The process of claim 19 wherein the vector is an H dimensional vector of token substrings and the index table entries also contain an attribute field, a position within the H dimensional vector and a frequency of reference records that map to the token substring contained in an index table entry.

34. A system for evaluating an input data record having fields containing data comprising:

a database for storing a reference table having a number of reference records against which an input data record is evaluated;

a preprocessor component for evaluating reference records in the reference table to identify tokens and determining a count of tokens in the reference table classified according to record field; said preprocessor evaluating reference records to identify tokens from said attribute fields and then evaluating each token to build a H dimensional vector of token substrings that represent the token; and building an index table wherein entries of the index table contains a token substring , an attribute field, a position within the H dimensional vector, and a list of reference records; and

a matching component for assigning a score to an input data record in relation to a reference record within the reference table by building a candidate record table of candidate records from the index table based on an H dimensional vector of token substrings determined from tokens contained in the input record and assigning a score to said candidate records based on a weight of the tokens of the input data record that is

based on a count of the tokens from the corresponding field contained within the reference table.

35. A data structure for use in evaluating an input data record having fields containing data comprising:

a reference table organized in attribute columns having a number of records against which an input data record is evaluated; and

an index table wherein each entry of the index table contains a token substring from a token in the reference table, a column of the reference table having said token from which the token substring is derived, a position within a H dimensional vector based on said token, and a list of records contained within the reference table.

36. The data structure of claim 35 wherein each entry of the index table additionally comprises an attribute field for the token from which a substring is derived.

37. A machine readable medium including instructions for evaluating an input data record having attribute fields containing by steps of:

accessing a reference table having a number of records organized into attribute fields against which an input data record is evaluated;

evaluating records in the reference table to identify tokens from said attribute fields and then evaluating each token with a function to build a vector of token substrings that serve as a signature of the token;

building an index table wherein each entry of the index table contains a token substring , a column of the reference table, a position within the vector, and a list of records contained within the reference table; and

looking up records in the index table based on the contents of the input record.

38. The machine readable medium of claim 37 additionally including instructions for implementing a step of assigning a similarity score to said input data record in relation to a candidate set of reference records within the reference table based on a combination of:

the number of common tokens of an evaluation field of the input data record and a corresponding field within a reference record from the reference table;

the similarity of the tokens that are not the same in the evaluation field of the input data record and the corresponding field of the reference record from the reference table; and

a weight of the tokens in the corresponding field of said reference record based on a count of the tokens from the corresponding field contained within the reference table.

39. The machine readable medium of claim 37 wherein a candidate record table is built and records from the index table are added to a candidate record table based on vector substring representations of the tokens of the input record.

40. The machine readable medium of claim 39 wherein a candidate record is added to the candidate record table only if a score assigned to the reference record can exceed a threshold based on an already evaluated substring representation of the input record.

41. The machine readable medium of claim 39 wherein once a likely reference record that matches the evaluation data record with a specified degree of certainty is found further searching for records in the reference table is stopped.

42. The machine readable medium of claim 38 wherein a closest K reference records from the reference table are identified as possible matches with the input record.

43. The machine readable medium of claim 38 wherein reference records having a similarity score greater than a threshold are identified as candidate records.

44. The machine readable medium of claim 38 additionally comprising the step of maintaining a token frequency cache in a high speed access memory for use in assigning weights to said tokens.

45. The machine readable medium of claim 38 wherein the tokens in different attribute fields are assigned different weights in determining said score.

46. The machine readable medium of claim 37 wherein the index table additionally comprises an attribute field for a token from which a substring is derived.